Fine-tuning Clair3 with 12 bacteria samples for improved variant calling performance

Drafted by Ruibang Luo, Zhenxian Zheng, Lei Chen from *the School of Computing and Data Science, HKU*

Acknowledgements: We thank Dr. William C. Shropshire from *The University of Texas MD Anderson Cancer Center* for providing the idea and testing datasets

May 19th, 2025

Fine-tuning Clair3 with 12 bacteria samples

Sample	Species	Biosample ID	SRA ID	Contigs	Used in training	Used for testing
ATCC_10708__202309	Salmonella enterica	SAMN38321309	SRR27638402	chromosome, plasmid		✓
ATCC_25922__202309	Escherichia coli	SAMN38321313	SRR27638398	chromosome, plasmid_1, plasmid_2, plasmid_3, plasmid_4		✓
ATCC_17802__202309	Vibrio parahaemolyticus	SAMN38321311	SRR27638400	chromosome_1, chromosome_2	✓	
ATCC_33560__202309	Campylobacter jejuni	SAMN38321314	SRR27638397	chromosome	✓	
ATCC_35221__202309	Campylobacter lari	SAMN38321315	SRR27638396	chromosome	✓	
ATCC_19119__202309	Listeria ivanovii	SAMN38321312	SRR27638399	chromosome	✓	
ATCC_35897__202309	Listeria welshimeri	SAMN38321316	SRR27638395	chromosome	✓	
ATCC_BAA-679__202309	Listeria monocytogenes	SAMN38321317	SRR27638394	chromosome	✓	
BPH2947__202310	Staphylococcus aureus	SAMN40453078	SRR28370694	chromosome, plasmid_1, plasmid_2	✓	
AJ292__202310	Klebsiella variicola	SAMN40453079	SRR28370693	chromosome	✓	
KPC2__202310	Klebsiella pneumoniae	SAMN40453080	SRR28370682	chromosome, plasmid_1, plasmid_2, plasmid_3	✓	
RDH275__202311	Streptococcus pyogenes	SAMN40453081	SRR28370671	chromosome	✓	
MMC234__202311	Streptococcus dysgalactiae	SAMN40453082	SRR28370660	chromosome	✓	
AMtb_1__202402	Mycobacterium tuberculosis	SAMN40453083	SRR28370649	chromosome	✓	

- Training dataset: 12 bacteria samples were used for model training.
- Testing datasets: 2 bacteria samples were held out to evaluate the fine-tuned Clair3 model performance.
- BAMs were subsampled 80%, 60%, 40%, and 20% read coverage for model training.
- The fine-tuning learning rate was initially set to 5e-7, and the max training epoch was set to 10. The epoch with best validation performance was selected for benchmark.

Hall M B, Wick R R, Judd L M, et al. Benchmarking reveals superiority of deep learning variant callers on bacterial nanopore sequence data[J]. Elife, 2024, 13: RP98300.

Performance of the fine-tuned Clair3 model on two holdout datasets

Caller	Sample	species	Variant type	Precision	Recall	F1-score	F1-Qscore	TRUTH TOTAL	TRUTH TP	TRUTH FN	QUERY TOTAL	QUERY TP	QUERY FP
Clair3 (v1.1.0)	ATCC_25922__202309	Escherichia coli	SNP	100.00%	98.06%	99.02%	20.09	4,531	4,443	88	4,441	4,441	0
Clair3 (v1.1.0)	ATCC_25922__202309	Escherichia coli	INDEL	100.00%	99.45%	99.72%	25.56	361	359	2	360	360	0
Clair3 (fine-tuned)	ATCC_25922__202309	Escherichia coli	SNP	100.00%	99.98%	99.99%	39.57	4,531	4,530	1	4,528	4,528	0
Clair3 (fine-tuned)	ATCC_25922__202309	Escherichia coli	INDEL	100.00%	100.00%	100.00%	100.00	361	361	0	362	362	0
Clair3 (v1.1.0)	ATCC_10708__202309	Salmonella enterica	SNP	100.00%	99.99%	100.00%	45.74	18,784	18,783	1	18,769	18,769	0
Clair3 (v1.1.0)	ATCC_10708__202309	Salmonella enterica	INDEL	99.75%	98.25%	98.99%	19.97	399	392	7	400	399	1
Clair3 (fine-tuned)	ATCC_10708__202309	Salmonella enterica	SNP	99.99%	99.95%	99.97%	35.33	18,784	18,774	10	18,773	18,772	1
Clair3 (fine-tuned)	ATCC_10708__202309	Salmonella enterica	INDEL	100.00%	98.75%	99.37%	22.00	399	394	5	395	395	0

- Performance evaluation using vcfdist (https://github.com/TimD1/vcfdist).
- Clair3 fine-tuned model performed better than Clair3 v1.1.0 model (using the r1041_e82_400bps_sup_v430 pretrained model) in Escherichia coli.
- Comparable SNP performance and better Indel performance in Salmonella enterica.

Result

- Command to run the fine-tuned Clair3 model:

```
docker run -it \
    -v ${INPUT_DIR}:${INPUT_DIR} \
    -v ${OUTPUT_DIR}:${OUTPUT_DIR} \
    hkubal/clair3:latest /opt/bin/run_clair3.sh \
      --bam_fn=${INPUT_DIR}/input.bam \
      --ref_fn=${INPUT_DIR}/ref.fa \
      --threads=${THREADS} \
      --platform=ont \
      --model_path=${INPUT_DIR}/r1041_e82_400bps_sup_v430_bacteria_finetuned” \
      --output=${OUTPUT_DIR} \
      --include_all_ctgs \
      --print_ref_calls \
      --haploid_precise \
      --no_phasing_for_fa \
      --enable_variant_calling_at_sequence_head_and_tail
```